Homework #4

See Canvas for HW #4 assignment due date.

A. Theoretical Problems

A.1 (12 points) Let $Y_1, \ldots, Y_n \overset{i}{\sim} Poisson(\lambda_i)$. Show that, if $\eta_i = \beta_0$, then the maximum likelihood estimator of λ_i is $\widehat{\lambda}_i = \bar{Y}$, for all $i = 1, \ldots, n$.

Hint: Write out the general log-likelihood for Poisson first using η_i and then follow through with plugging in the value and evaluating/solving.

B. Computational Problems

Problem B.1

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study of 768 adult female Pima Indians living near Phoenix, AZ. The purpose of the study was the investigate factors related to diabetes.

(a) Perform simple graphical and numerical summaries of the data. Can you find any obvious irregularities in the data? If so, take appropriate steps to correct these problems.

```
# Find the data here..

pima = read.table("https://www.colorado.edu/amath/sites/default/files/attached-files/pima.txt",
                  sep = "\t", header = TRUE)

#Here's a description of the data: https://rdrr.io/cran/faraway/man/pima.html
```

(b) Fit a model with the result of the diabetes test as the response and all the other variables as predictors. Store this model as `glmod_pima`. Can you tell whether this model fits the data?

(c) Using the model above, write R code to calculate the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming all other factors are held constant. Store your answer in a variable `x`.

Also, give a confidence interval for this difference, stored in a variable `ci`.

(d) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

(e) Ethical Issues in Data Collection

Read Maya Iskandarani's piece (https://researchblog.duke.edu/2016/10/24/diabetes-and-privacy-meet-big-data/)on consent and privacy concerns raised by this dataset. Summarize those concerns here.

Problem B.2

The ships dataset (in the MASS package) gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation.

(a) The code below splits the data into a training set (80% of the data) and a test set (the remaining 20%). Use the training set to develop an appropriate regression model for the rate of incidents, using type, period, and year as predictors (HINT: is this a count model or a rate model?). Store this model in `glmod_ships`.

```
library(MASS)
data(ships)
ships = ships[ships$service != 0,]
ships$year = as.factor(ships$year)
ships$period = as.factor(ships$period)


dim(ships)
set.seed(11)
n = floor(0.8 * nrow(ships))
index = sample(seq_len(nrow(ships)), size = n)

train = ships[index, ]
test = ships[-index, ]
head(train)
dim(train)
summary(train)
```

(b) Use the model that you stored in glmod_ships to calculate the mean squared prediction error (MSPE) for the test set. Store the predicted MSPE in `mse_glmod_ships`.

Recall from earlier assignments that the MSE can give us a sense of how well the model does at predicting new observations. The predicted mean squared error (MSE) is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2,$$

where y_i is the response in the test set, and $\hat{y}_i$ is the predicted response from glmod_ships, given the predictor values in the test set.

Note that the predict.glm() function can be helpful here. Just be sure to specify the type argument (HINT: do you want $\hat{y}_i$ to be on the scale of the linear predictor η, or the mean of the response?)

(c) Now construct a new regression model leaving out the year predictor. Store this model as `glmod_ships2`. Calculate the predicted MSPE (Mean Squared Prediction Error) for the test set using `glmod_ships2`. Decide which model is better - `glmod_ships` or `glmod_ships2` - and store your answer in `glmod_ships3`.

(d) Let $\alpha = 0.05$. Conduct two χ^2 tests (using the deviance):

1. Test the adequacy of null model (store the p-value for this test in chisq_null); and

2. Test the adequacy of the glmod_ships model agaisnt the saturated model (store the p-value for this test in chisq_p).

What conclusions should you draw from these tests?

(e) Plot the deviance residuals against the linear predictor η. Interpret this plot. Hint: The residuals function has a type parameter and "deviance" is one possible type.

(f) For some GLMs (including the type in this question!), *overdispersion* is sometimes a problem. *Overdispersion* occurs when the observed (data) variance is higher than expected, if the model is correct. Explore the two models above for evidence of overdispersion.

```
library(AER)
#this package has a function overdispersiontest(), which conducts an overdisperion test.
#If you use it, please clearly describe the test being used, including hypotheses, test statistic distribution,
#and conclusions
```